Cross-Modal Collaborative Representation Learning and a Large-Scale RGBT Benchmark for Crowd Counting

Lingbo Liu[1], Jiaqi Chen[1], Hefeng Wu[1], Guanbin Li[1,2], Chenglong Li[3], Liang Lin[1,4*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Pazhou Lab, Guangzhou, China [3] Anhui University, China [4] DarkMatter AI Research, China

Abstract

Crowd counting is a fundamental yet challenging task, which desires rich information to generate pixel-wise crowd density maps. However, most previous methods only used the limited information of RGB images and cannot well discover potential pedestrians in unconstrained scenarios. In this work, we find that incorporating optical and thermal information can greatly help to recognize pedestrians. To promote future researches in this field, we introduce a large-scale RGBT Crowd Counting (RGBT-CC) benchmark, which contains 2,030 pairs of RGB-thermal images with 138,389 annotated people. Furthermore, to facilitate the multimodal crowd counting, we propose a cross-modal collaborative representation learning framework, which consists of multiple modality-specific branches, a modality-shared branch, and an Information Aggregation-Distribution Module (IADM) to capture the complementary information of different modalities fully. Specifically, our IADM incorporates two collaborative information transfers to dynamically enhance the modality-shared and modality-specific representations with a dual information propagation mechanism. Extensive experiments conducted on the RGBT-CC benchmark demonstrate the effectiveness of our framework for RGBT crowd counting. Moreover, the proposed approach is universal for multimodal crowd counting and is also capable to achieve superior performance on the ShanghaiTechRGBD [22] dataset. Finally, our source code and benchmark have been released at http://lingboliu.com/RGBT_Crowd_Counting.html.

1. Introduction

Crowd counting [18, 10] is a fundamental computer vision task that aims to automatically estimate the number of people in unconstrained scenes. Over the past decade, this task has attracted a lot of research interests due to its huge application potentials (e.g., traffic management [62, 28] and video surveillance [52]). During the recent COVID-19 pandemic [47], crowd counting has also been employed widely for social distancing monitoring [11].

In the literature, numerous models [64, 43, 27, 56, 1, 21, 26, 34, 30, 32] have been proposed for crowd counting. Despite substantial progress, it remains a very challenging problem that desires rich information to generate pixel-wise crowd density maps. However, most previous methods only utilized the optical information extracted from RGB images and may fail to accurately recognize the semantic objects in unconstraint scenarios. For instance, as shown in Fig. 1-(a,b), pedestrians are almost invisible in poor illumination conditions (such as backlight and night) and they are hard to be directly detected from RGB images. Moreover, some human-shaped objects (e.g., tiny pillars and blurry traffic lights) have similar appearances to pedestrians [59] and they are easily mistaken for people when relying solely on optical features. In general, RGB images cannot guarantee the high-quality density maps, and more comprehensive information should be explored for crowd counting.

Fortunately, we observe that thermal images can greatly facilitate distinguishing the potential pedestrians from cluttered backgrounds. Recently, thermal cameras have been extensively popularized due to the COVID-19 pandemic, which increases the feasibility of thermal-based crowd counting. However, thermal images are not perfect. As shown in Fig. 1-(c,d), some hard negative objects (e.g., heating walls and lamps) are also highlighted in thermal images, but they can be eliminated effectively with the aid of optical information. Overall, RGB images and thermal images are highly complementary. To the best of our knowledge, no attempts have been made to simultaneously explore RGB and thermal images for estimating the crowd counts. In this work, to promote further researches of this field, we propose a large-scale benchmark "RGBT Crowd Counting (**RGBT-CC**)", which contains 2,030 pairs of RGB-thermal images and 138,389 annotated pedestrians. Moreover, our benchmark makes significant advances in terms of diversity and difficulty, as these RGBT images were captured from un-

[*] The corresponding author is Liang Lin. Lingbo Liu and Jiaqi Chen share first-authorship.

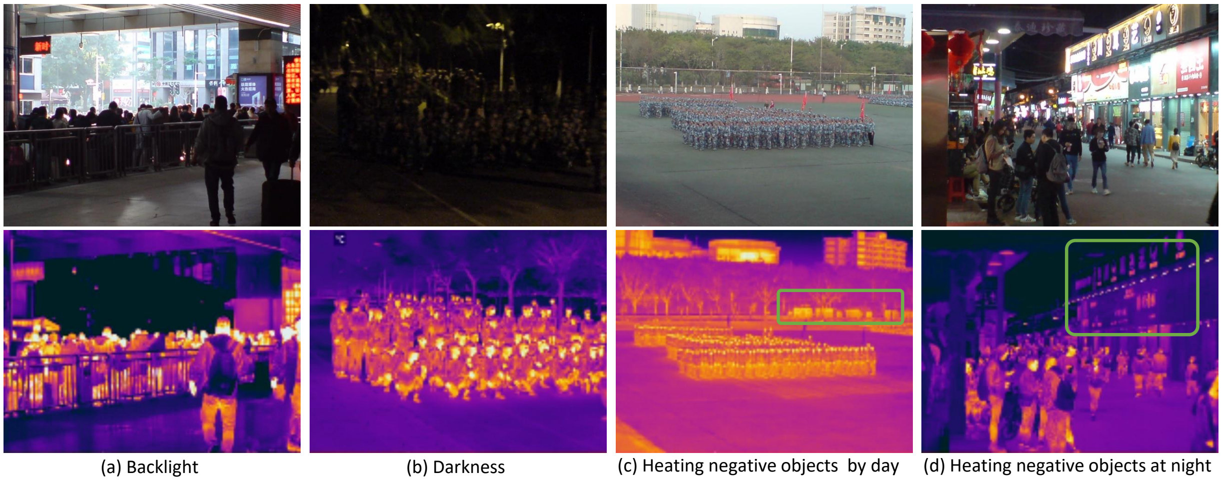

(a) Backlight
(b) Darkness
(c) Heating negative objects by day
(d) Heating negative objects at night

Figure 1. Visualization of RGB-thermal images in our RGBT-CC benchmark. When only using optical information of RGB images, we cannot effectively recognize pedestrians in poor illumination conditions, as shown in (a) and (b). When only utilizing thermal images, some heating negative objects are hard to be distinguished, as shown in (c) and (d).

constrained scenes (e.g., malls, streets, train stations, etc.) with various illumination (e.g., day and night).

Nevertheless, capturing the complementarities of multimodal data (i.e., RGB and thermal images) is non-trivial. Conventional methods [22, 67, 37, 15, 54, 46] either feed the combination of multimodal data into deep neural networks or directly fuse their features, which could not well exploit the complementary information. In this work, to facilitate the multimodal crowd counting, we introduce a cross-modal collaborative representation learning framework, which incorporates multiple modality-specific branches, a modality-shared branch, and an Information Aggregation-Distribution Module (IADM) to fully capture the complementarities among different modalities. Specifically, our IADM is integrated with two collaborative components, including **i)** an Information Aggregation Transfer that dynamically aggregates the contextual information of all modality-specific features to enhance the modality-shared feature and **ii)** an Information Distribution Transfer that propagates the modality-shared information to symmetrically refine every modality-specific feature for further representation learning. Furthermore, the tailor-designed IADM is embedded in different layers to learn the cross-modal representation hierarchically. Consequently, the proposed framework can generate knowledgeable features with comprehensive information, thereby yielding high-quality crowd density maps.

It is worth noting that our method has three appealing properties. **First**, thanks to the dual information propagation mechanism, IADM can effectively capture the multimodal complementarities to facilitate the crowd counting task. **Second**, as a plug-and-play module, IADM can be easily incorporated into various backbone networks for end-to-end optimization. **Third**, our framework is universal for multimodal crowd counting. Except for RGBT counting, the proposed method can also be directly applied for RGB-Depth counting. In summary, the major contributions of this work are three-fold:

- We introduce a large-scale RGBT benchmark to promote the research of crowd counting, in which 138,389 pedestrians are annotated in 2,030 pairs of RGB-thermal images captured in unconstrained scenarios.
- We develop a cross-modal collaborative representation learning framework, which is capable of fully learning the complementarities among different modalities with a Information Aggregation-Distribution Module.
- Extensive experiments conducted on two multimodal benchmarks (i.e., RGBT-CC and ShanghaiTechRGBD [22]) greatly demonstrate that the proposed method is effective and universal for multimodal crowd counting.

2. Related Works

Crowd Counting Benchmarks: In recent years, we have witnessed the rapid evolution of crowd counting benchmarks. UCSD [3] and WorldExpo [57] are two early datasets that respectively contain 2,000 and 3,980 video frames with low diversities and low-medium densities. To alleviate the limitations of the aforementioned datasets, Zhang *et al.* [64] collected 1,198 images with 330,165 annotated heads, which are of better diversity in terms of scenes and density levels. Subsequently, three large-scale datasets were proposed in succession. For instance, UCF-QNRF [14] is composed of 1,535 high density images images with a total of 1.25 million pedestrians. JHU-CROWD++ [45] contains 4,372 images with 1.51 million annotated heads, while NWPU-Crowd [50] consists of 2.13 million annotations in 5,109 images. Nevertheless, all the above benchmarks are based on RGB optical images, in which almost all previous methods fail to recognize the invisible pedestrians in poor illumination conditions. Recently, Lian *et al.* [22] utilized a stereo camera to capture 2,193 depth images that are insensitive to illumination. However, these images are coarse in outdoor scenes due to the limited depth ranges (0∼20 meters), which seriously restricts their deployment scopes. Fortunately, we find that thermal images are robust to illumination and have large perception distance, thus can help to recognize pedestrians under various scenarios. Therefore, we propose the first RGBT crowd counting dataset in this work, hoping that it would greatly promote the future development in this field.

Crowd Counting Approaches: As a classics problem in computer vision, crowd counting has been studied extensively. Early works [4, 5, 13] directly predict the crowd count with regression models, while subsequent methods usually generate crowd density maps and then accumulate all pixels' values to obtain the final counts. Specifically, a large number of deep neural networks with various architectures [9, 57, 49, 48, 41, 17, 43, 21, 55, 29, 39, 16, 53] and loss functions [2, 14, 34, 26] are developed for still image-based crowd counting. Meanwhile, some methods [60, 52, 40, 31] perform crowd estimation from multi-view images or surveillance videos. However, all aforementioned methods estimate crowd counts with the optical information of RGB images/videos and are not effective when working in poor illumination conditions. Recently, depth images are used as auxiliary information to count and locate human heads [22]. Nevertheless, depth images are coarse in outdoor scenarios, thus depth-based methods have relatively limited deployment scopes. Nevertheless, depth images are coarse in outdoor scenarios, thus depth-based methods have relatively limited deployment scopes.

Multi-Modal Representation Learning: Multi-modal representation learning aims at comprehending and representing cross-modal data through machine learning. There are many strategies in cross-modal feature fusion. Some simple fusion methods [19, 22, 46, 8] obtain a fused feature with the operations of element-wise multiplication/addition or concatenation in the "Early Fusion" and "Late Fusion" way. To exploit the advantages of both early and late fusion, various two-stream-based models [51, 38, 66, 63] propose to fuse hierarchical cross-modal features, achieving the fully representative shared feature. Besides, a few approaches [33] explore the use of a shared branch, mapping the shared information to common feature spaces. Furthermore, some recent works [7, 35, 58] are presented to address RGBD saliency detection, which is also a cross-modal dense prediction task like RGBT crowd counting. However, most of these works are one-way information transfer, just using depth modality as auxiliary information to help the representation learning of RGB modality. In this work, we propose a symmetric dynamic enhancement mechanism that can take full advantage of the modal complementarities in crowd counting.

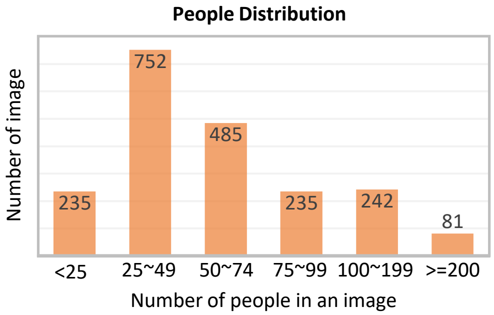


Figure 2. The statistics histogram of people distribution in the proposed RGBT Crowd Counting benchmark.

Table 1. The training, validation and testing sets of our RGBT-CC benchmark. In each grid, the first value is the number of images, while the second value denotes the average count per image.

	Training	Validation	Testing
#Bright	510 / 65.66	97 / 63.02	406 / 73.39
#Dark	520 / 62.52	103 / 67.74	394 / 74.88
#Total	1030 / 64.07	200 / 65.45	800 / 74.12
Scene	malls, streets, train/metro stations, etc		

3. RGBT Crowd Counting Benchmark

To the best of our knowledge, there is currently no public RGBT dataset for crowd counting. To promote the future research of this task, we propose a large-scale RGBT Crowd Counting (**RGBT-CC**) benchmark. Specifically, we first use an optical-thermal camera to take a large number of RGB-thermal images in various scenarios (e.g., malls, streets, playgrounds, train stations, metro stations, etc). Due to the different types of electronic sensors, original RGB images have a high resolution of 2,048×1,536 with a wider field of view, while thermal images have a standard resolution of 640×480 with a smaller field of view. On the basis of coordinate mapping relation, we crop the corresponding RGB regions and resize them to 640×480. We then choose 2,030 pairs of representative RGB-thermal images for manual annotations. Among these samples, 1,013 pairs are captured in the light and 1,017 pairs are in the darkness. A total of 138,389 pedestrians are marked with point annotations, on average 68 people per image. The detailed distribution of people is shown in Fig. 2. Finally, the proposed RGBT-CC benchmark is randomly divided into three parts. As shown in Table 1, 1030 pairs are used for training, 200 pairs are for validation and 800 pairs are for testing. Compared with those Internet-based datasets [14, 50, 45] with serious bias, our RGBT-CC dataset has closer crowd density distribution to realistic cities, since our images are captured in urban scenes with various densities. Therefore, our dataset has wider applications for urban crowd analysis.

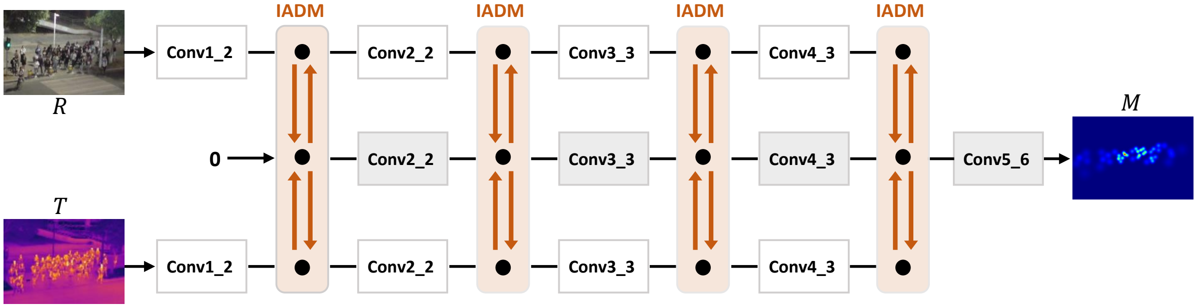


Figure 3. The architecture of the proposed cross-modal collaborative representation learning framework for multimodal crowd counting. Specifically, our framework consists of multiple modality-specific branches, a modality-shared branch, and an Information Aggregation-Distribution Module (IADM).

4. Method

In this work, we propose a cross-modal collaborative representation learning framework for multimodal crowd counting. Specifically, multiple modality-specific branches, a modality-shared branch, and an Information Aggregation-Distribution Module are incorporated to fully capture the complementarities among different modalities with a dual information propagation paradigm. In this section, we adopt the representative CSRNet [21] as a backbone network to develop our framework for RGBT crowd counting. It is worth noting that our framework can be implemented with various backbone networks (e.g., MCNN [64], SANet [2], and BL [34]), and is also universal for multimodal crowd counting, as verified in Section 5.4 by directly applying it to the ShanghaiTechRGBD [22] dataset.

4.1. Overview

As shown in Fig. 3, the proposed RGBT crowd counting framework is composed of three parallel backbones and an Information Aggregation-Distribution Module (IADM). Specifically, the top and bottom backbones are developed for modality-specific (i.e. RGB images and thermal images) representation learning, while the middle backbone is designed for modality-shared representation learning. To fully exploit the multimodal complementarities, our IADM dynamically transfers the specific-shared information to collaboratively enhance the modality-specific and modality-shared representations. Consequently, the final modality-shared feature contains comprehensive information and facilitates generating high-quality crowd density maps.

Given an RGB image R and a thermal image T, we first feed them into different branches to extract modality-specific features, which maintain the specific information of individual modality. The modality-shared branch takes a zero-tensor as input[1] and aggregates the information of modality-specific features hierarchically. As mentioned above, each branch is implemented with CSRNet, which consists of (1) a front-end block with the first ten convolutional layers of VGG16 [42] and (2) a back-end block with six dilated convolutional layers. More specifically, the modality-specific branches are based on the CSRNet front-end block, while the modality-shared branch is based on the last 14 convolutional layers of CSRNet. In our work, the j-th dilated convolutional layer of back-end block is renamed as "Conv5_j". For convenience, the RGB, thermal, and modality-shared features at Conv$i$_j layer are denoted as $F_r^{i,j}$, $F_t^{i,j}$, and $F_s^{i,j}$, respectively.

After feature extraction, we employ the Information Aggregation-Distribution Module described in Section 4.2 to learn cross-modal collaborative representation. To exploit the multimodal information hierarchically, the proposed IADM is embedded after different layers, such as Conv1_2, Conv2_2, Conv3_3, and Conv4_3. Specifically, after Conv$i$_j, IADM dynamically transfers complementary information among modality-specific and modality-shared features for mutual enhancement. This process can be formulated as follow:

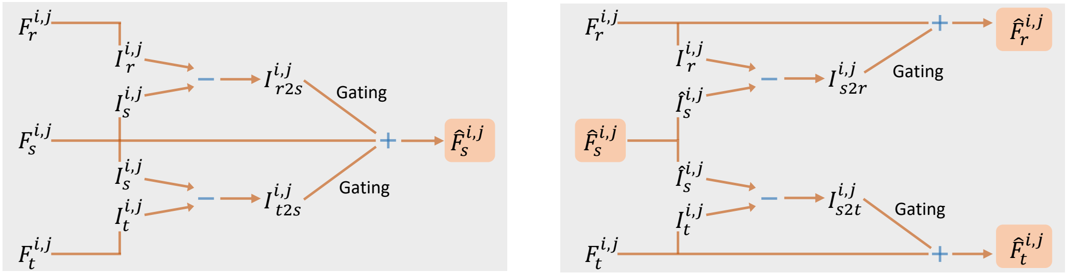
$$\hat{F}_s^{i,j}, \hat{F}_r^{i,j}, \hat{F}_t^{i,j} = \mathrm{IADM}(F_s^{i,j}, F_r^{i,j}, F_t^{i,j}), \tag{1}$$

where $\hat{F}_s^{i,j}$, $\hat{F}_r^{i,j}$, and $\hat{F}_t^{i,j}$ are the enhanced features of $F_s^{i,j}$, $F_r^{i,j}$, and $F_t^{i,j}$ respectively. These features are then fed into the next layer of each branch to further learn high-level multimodal representations. Thanks to the tailor-designed IADM, the complementary information of the input RGB image and the thermal image is progressively transferred into the modality-shared representations. The final modality-shared feature $F_s^{5,6}$ contains rich information. Finally, we directly feed $F_s^{5,6}$ into a 1*1 convolutional layer for prediction of the crowd density map M.

4.2. Collaborative Representation Learning

As analyzed in Section , RGB images and thermal images are highly complementary. To fully capture their complementarities, we propose an Information Aggregation and Distribution Module (IADM) to collaboratively learn cross-

[1] When the input of modality-shared branch is set to 0, Eq.3 at Conv1_2 is simplified as $\hat{F}_s^{1,2} = I_r^{1,2} \odot Conv_{1*1}(I_r^{1,2}) + I_t^{1,2} \odot Conv_{1*1}(I_t^{1,2})$. In other words, the initial modality-shared feature is generated by directly aggregating the information of RGB and thermal features.



(a) Information Aggregation

(b) Information Distribution

Figure 4. **(a) Information Aggregation Transfer:** we first extract the contextual information I_r/I_t from modality-specific features F_r/F_t, and then propagate them dynamically to enhance the modality-shared feature F_s. **(b) Information Distribution Transfer:** the contextual information $\hat{I}_s$ of the enhance feature $\hat{F}_s$ is distributed adaptively to each modality-specific feature for feedback refinement. "+" denotes element-wise addition and "-" refers to element-wise subtraction.

modal representation with a dual information propagation mechanism. Specifically, our IADM is integrated with two collaborative transfers, which dynamically propagate the contextual information to mutually enhance the modality-specific and modality-shared representations.

1) Contextual Information Extraction: In this module, we propagate the contextual information rather than the original features, because the later manner would cause the excessive mixing of specific-shared features. To this end, we employ a L-level pyramid pooling layer to extract the contextual information for a given feature $F^{i,j} \in R^{h\times w\times c}$. Specifically, at the l^{th} level (l=1,...,L), we apply a $2^{l-1}\times 2^{l-1}$ max-pooling layer to generate a $\frac{h}{2^{l-1}}\times\frac{w}{2^{l-1}}$ feature, which is then upsampled to $h\times w$ with nearest neighbor interpolation. For convenience, the upsampled feature is denoted as $F^{i,j,l}$. Finally, the contextual information $I^{i,j} \in R^{h\times w\times c}$ of feature $F^{i,j}$ is computed as:

$$I^{i,j} = Conv_{1*1}(F^{i,j,1} \oplus F^{i,j,2} \oplus ... \oplus F^{i,j,L}), \qquad (2)$$

where $\oplus$ denotes an operation of feature concatenation and $Conv_{1*1}$ is a 1*1 convolutional layer. This extraction has two advantages. First, with a larger receptive field, each position at $I^{i,j}$ contains more context. Second, captured by different sensors, RGB images and thermal images are not strictly aligned, as shown in Figure 1. Thanks to the translation invariance of max-pooling layers, we can eliminate the misalignment of RGB-thermal images to some extent.

2) Information Aggregation Transfer (IAT): In our work, IAT is proposed to aggregate the contextual information of all modality-specific features to enhance the modality-shared feature. As shown in Fig. 4-(a), instead of directly absorbing all information, our IAT transfers the complementary information dynamically with a gating mechanism that adaptively filters useful information. Specifically, given features $F_r^{i,j}$, $F_t^{i,j}$ and $F_s^{i,j}$, we first extract their contextual information $I_r^{i,j}$, $I_t^{i,j}$, and $I_s^{i,j}$ with Eq. 2. Similar to [61, 65], we then obtain two residual information $I_{r2s}^{i,j}$ and $I_{t2s}^{i,j}$ by computing the differences between $I_r^{i,j}/I_t^{i,j}$ and $I_s^{i,j}$. Finally, we apply two gating functions to adaptively propagate the complementary information for refining the modality-shared feature $F_s^{i,j}$. The enhanced feature $\hat{F}_s^{i,j}$ is formulated as follow:

$$\begin{aligned} I_{r2s}^{i,j} &= I_r^{i,j} - I_s^{i,j}, \;\; w_{r2s}^{i,j} = Conv_{1*1}(I_{r2s}^{i,j}), \\ I_{t2s}^{i,j} &= I_t^{i,j} - I_s^{i,j}, \;\; w_{t2s}^{i,j} = Conv_{1*1}(I_{t2s}^{i,j}), \\ \hat{F}_s^{i,j} &= F_s^{i,j} + I_{r2s}^{i,j}\odot w_{r2s}^{i,j} + I_{t2s}^{i,j}\odot w_{t2s}^{i,j}, \end{aligned} \qquad (3)$$

where the gating functions are implemented by convolutional layers, $w_{r2s}^{i,j}$ and $w_{t2s}^{i,j}$ are the gating weights. $\odot$ refers to an operation of element-wise multiplication. With such a mechanism, the complementary information is effectively embedded into the modality-shared representation, thus our method can better exploit the multimodal data.

3) Information Distribution Transfer (IDT): After information aggregation, we distribute the information of the new modality-shared feature to refine each modality-specific feature respectively. As shown in Fig. 4-(b), with the enhanced feature $\hat{F}_s^{i,j}$, we first extract its contextual information $\hat{I}_s^{i,j}$, which is then dynamically propagated to $F_r^{i,j}$ and $F_t^{i,j}$. Simialr to IAT, two gating functions are used for information filtering. Specifically, the enhanced modality-specific features are computed as follow:

$$\begin{aligned} I_{s2r}^{i,j} &= \hat{I}_s^{i,j} - I_r^{i,j}, & I_{s2t}^{i,j} &= \hat{I}_s^{i,j} - I_t^{i,j}, \\ w_{s2r}^{i,j} &= Conv_{1*1}(I_{s2r}^{i,j}), & w_{s2t}^{i,j} &= Conv_{1*1}(I_{s2t}^{i,j}), \\ \hat{F}_r^{i,j} &= F_r^{i,j} + I_{s2r}^{i,j}\odot w_{s2r}^{i,j}, & \hat{F}_t^{i,j} &= F_t^{i,j} + I_{s2t}^{i,j}\odot w_{s2t}^{i,j}. \end{aligned}$$

Finally, all enhanced features $\hat{F}_r^{i,j}$, $\hat{F}_t^{i,j}$, and $\hat{F}_s^{i,j}$ are fed into the following layers of the individual branch for further representation learning.

5. Experiments

5.1. Implementation Details & Evaluation Metrics

In this work, the proposed method is implemented with PyTorch [36]. Here we take various models (e.g., CSRNet [21], MCNN [64], SANet [2], and BL [34]) as backbone to develop multiple instances of our framework. To maintain a similar number of parameters to original mod-

Table 2. The performance of different inputs and different representation learning approaches on our RGBT-CC benchmark.

Input Data	Representation Learning	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
RGB	-	33.94	40.76	47.31	57.20	69.59
T	-	21.64	26.22	31.65	38.66	37.38
RGBT	Early Fusion	20.40	23.58	28.03	35.51	35.26
	Late fusion	19.87	25.60	31.93	41.60	35.09
	W/O Gating Mechanism	19.76	23.60	28.66	36.21	33.61
	W/O Modality-Shared Feature	18.67	22.67	27.95	36.04	33.73
	W/O Information Distribution	18.59	23.08	28.73	36.74	32.91
	IADM	**17.94**	**21.44**	**26.17**	**33.33**	**30.91**

Table 3. The performance under different illumination conditions on our RGBT-CC benchmark. The unimodal data is directly fed into CSRNet, while the multimodal data is fed into our proposed framework based on CSRNet.

Illumination	Input Data	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
Brightness	RGB	23.49	30.14	37.47	48.46	45.40
	T	25.21	28.98	34.82	42.25	40.60
	RGBT	**20.36**	**23.57**	**28.49**	**36.29**	**32.57**
Darkness	RGB	44.72	51.70	57.45	66.21	87.81
	T	17.97	23.38	28.39	34.95	33.74
	RGBT	**15.44**	**19.23**	**23.79**	**30.28**	**29.11**

els for fair comparisons, the channel number of these backbones in our framework is respectively set to 70%, 60%, 60%, and 60% of their original values. The kernel parameters are initialized by Gaussian distribution with a zero mean and a standard deviation of 1e-2. At each iteration, a pair of 640×480 RGBT image is fed into the network. The ground-truth density map is generated with geometry-adaptive Gaussian kernels [64]. The learning rate is set to 1e-5 and Adam [20] is used to optimize our framework. Notice that the loss function of our framework is the same as that of the adopted backbone network.

Following [25, 44, 24], we adopt the Root Mean Square Error (RMSE) as an evaluation metric. Moreover, Grid Average Mean Absolute Error (GAME [12]) is utilized to evaluate the performance in different regions. Specifically, for a specific level l, we divide the given images into 4^l non-overlapping regions and measure the counting error in each region. Finally, the GAME at level l is computed as:

$$GAME(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l}|\hat{P}_i^j - P_i^j|, \qquad (4)$$

where N is the total number of the testing samples, $\hat{P}_i^j$ and P_i^j are the estimated count and the corresponding ground-truth count in the j^{th} region of the i^{th} image. Note that GAME(0) is equivalent to Mean Absolute Error (MAE).

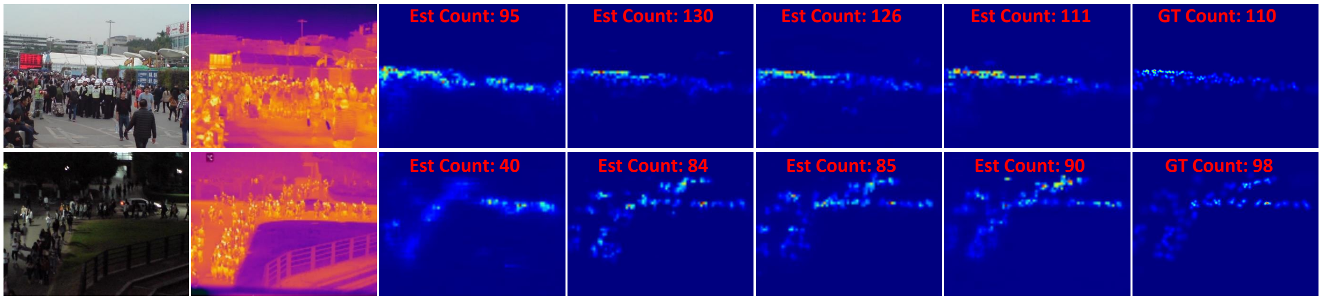
5.2. Ablation Studies

We perform extensive ablation studies to verify the effectiveness of each component in our framework. In this subsection, CSRNet is utilized as the backbone network to implement our proposed method.

1) Effectiveness of Multimodal Data: We first explore whether the multimodal data (i.e., RGB images and thermal images) is effective for crowd counting. As shown in Table 2, when only feeding RGB images into CSRNet, we obtain less impressive performance (e.g., GAME(0) is 33.94 and RMSE is 69.59), because we cannot effectively recognize people in dark environments. When utilizing thermal images, GAME(0) and RMSE are sharply reduced to 21.64 and 37.38, which demonstrates that thermal images are more useful than RGB images. In contrast, various models in the bottom six rows of Table 2 achieve better performance, when considering RGB and thermal images simultaneously. In particular, our CSRNet+IADM has a relative performance improvement of 17.3% on RMSE, compared with the thermal-based CSRNet.

To further verify the complementarities of multimodal data, the testing set is divided into two parts to measure the performance in different illumination conditions separately. As shown in Table 3, using both RGB and thermal images, our CSRNet+IADM consistently outperforms the unimodal CSRNet in both bright and dark scenarios. This is attributed to the thermal information that greatly helps to distinguish potential pedestrians from the cluttered background, while optical information is beneficial to eliminate heating negative objects in thermal images. Moreover, we also visualize some crowd density maps generated with different modal data in Fig. 4. We can observe that the density maps and estimated counts of our CSRNet+IADM are more accurate. These quantitative and qualitative experiments show that RGBT images are greatly effective for crowd counting.

2) Which Representation Learning Method is Better? We implement six methods for multimodal representation learning. Specifically, “Early Fusion” feeds the concatenation of RGB and thermal images into CSRNet. “Late Fusion” extracts the RGB and thermal features respectively



(a) RGB images (b) Thermal images (c) RGB results (d) Thermal results (e) Early fusion results (f) Our results (g) Ground-truth

Figure 5. Visualization of the crowd density maps generated in different illumination conditions. (a) and (b) show the input RGB images and thermal images. (c) and (d) are the results of RGB-based CSRNet and thermal-based CSRNet. (e) shows the results of CSRNet that takes the concatenation of RGB and thermal images as input. (f) refers to the results of our CSRNet+IDAM. And the ground-truths are shown in (g). We can observe that our density maps and estimated counts are more accurate than those of other methods. (*Best to zoom in to view this figure.*)

Table 4. Performance of different methods on the proposed RGBT-CC benchmark. All the methods in this table utilize both RGB images and thermal images to estimate the crowd counts.

Backbone	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
UCNet [58]	33.96	42.42	53.06	65.07	56.31
HDFNet [35]	22.36	27.79	33.68	42.48	33.93
BBSNet [7]	19.56	25.07	31.25	39.24	32.48
MVMS [60]	19.97	25.10	31.02	38.91	33.97
MCNN	21.89	25.70	30.22	37.19	37.44
MCNN + IADM	**19.77**	**23.80**	**28.58**	**35.11**	**30.34**
SANet	21.99	24.76	28.52	34.25	41.60
SANet + IADM	**18.18**	**21.84**	**26.27**	**32.95**	**33.72**
CSRNet	20.40	23.58	28.03	35.51	35.26
CSRNet + IADM	**17.94**	**21.44**	**26.17**	**33.33**	**30.91**
BL	18.70	22.55	26.83	34.62	32.67
BL + IADM	**15.61**	**19.95**	**24.69**	**32.89**	**28.18**

Table 5. Performance of different level numbers of the pyramid pooling layer in IADM.

#Level	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
L=1	18.94	23.05	28.03	35.88	33.01
L=2	18.35	22.56	27.84	35.90	31.94
L=3	**17.94**	**21.44**	**26.17**	**33.33**	**30.91**
L=4	17.80	21.39	25.91	33.20	31.48

with two CSRNet and then combines their features to generate density maps. As shown in Table 2, these two models are better than unimodal models, but their performance still lags far behind various variants of our IADM. For instance, without gating functions, the variant "W/O Gating Mechanism" directly propagates information among different features and obtains an RMSE of 33.61. The variant "W/O Modality-Shared Feature" obtains a GAME(0) of 18.67 and an RMSE of 33.73, when removing the modality-shared branch and directly refining the modality-specific features. When using the modality-shared branch but only aggregating multimodal information, the variant "W/O Information Distribution" obtains a relatively better RMSE 32.91. When using the full IADM, our method achieves the best performance on all evaluation metrics. This is attributed to our tailor-designed architecture (i.e., specific-shared branches, dual information propagation) that can effectively learn the multimodal collaborative representation, and fully capture the complementary information of RGB and thermal images. These experiments demonstrate the effectiveness of the proposed IADM for multimodal representation learning.

3) The Effectiveness of Level Number of Pyramid Pooling Layer: In the proposed IDAM, an L-level pyramid pooling layer is utilized to extract contextual information. In this section, we explore the effectiveness of the level number. As shown in Table 5, when L is set to 1, the GAME(3) and RMSE are 35.88 and 33.01, respectively. As the level number increases, our performance also becomes better gradually, and we can achieve very competitive results when the pyramid pooling layer has three levels. More levels over 3 will not bring additional performance gains. Thus, the level number L is consistently set to 3 in our work.

5.3. Comparison with State-of-the-Art Methods

We compare the proposed method with state-of-the-art methods on the large-scale RGBT-CC benchmark. The compared methods can be divided into two categories. The first class is the specially-designed models for crowd counting, including MCNN [64], SANet [2], CSRNet [21], and

Table 6. Performance of different methods on the ShanghaiTechRGBD benchmark. All the methods in this table utilize both RGB images and depth images to estimate the crowd counts.

Method	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
UCNet [58]	10.81	15.24	22.04	32.98	15.70
HDFNet [35]	8.32	13.93	17.97	22.62	13.01
BBSNet [7]	6.26	8.53	11.80	16.46	9.26
DetNet [23]	9.74	-	-	-	13.14
CL [14]	7.32	-	-	-	10.48
RDNet [22]	4.96	-	-	-	7.22
MCNN	11.12	14.53	18.68	24.49	16.49
MCNN + IADM	**9.61**	**11.89**	**15.44**	**20.69**	**14.52**
BL	8.94	11.57	15.68	22.49	12.49
BL + IADM	**7.13**	**9.28**	**13.00**	**19.53**	**10.27**
SANet	5.74	7.84	10.47	14.30	8.66
SANet + IADM	**4.71**	**6.49**	**9.02**	**12.41**	**7.35**
CSRNet	4.92	6.78	9.47	13.06	7.41
CSRNet + IADM	**4.38**	**5.95**	**8.02**	**11.02**	**7.06**

BL [34]. These methods are reimplemented to take the concatenation of RGB and thermal images as input in an "Early Fusion" way. Moreover, MVMS [60] is also reimplemented on RGBT-CC and pixel-wise attention map [6] is utilized to fuse the features of optical view and thermal view. The second class is several best-performing models for multimodal learning, including UCNet [58], HDFNet [35], and BBSNet [7]. Based on their official codes, these methods are reimplemented to estimate crowd counts on our RGBT-CC dataset. As mentioned above, our IADM can be incorporated into various networks, thus here we take CSRNet, MCNN, SANet, and BL as backbone to develop multiple instances of our framework.

The performance of all comparison methods is shown in Table 4. As can be observed, all instances of our method outperform the corresponding backbone networks consistently. For instance, both MCNN+IADM and SANet+IADM have a relative performance improvement of 18.9% on RMSE, compared with their "Early Fusion" models. Moreover, our CSRNet+IADM and BL+IADM achieve better performance on all evaluation metrics, compared with other advanced methods (i.e., UCNet, HDFNet, and BBSNet). This is because our method learns specific-shared representations explicitly and enhances them mutually, while others just simply fuse multimodal features without mutual enhancements. Thus our method can better capture the complementarities of RGB images and thermal images. This comparison has demonstrated the effectiveness of our method for RGBT crowd counting.

5.4. Apply to RGBD Crowd Counting

We apply the proposed method to estimate crowd counts from RGB images and depth images. In this subsection, we also take various crowd counting models as backbone to develop our framework on ShanghaiTechRGBD [22] benchmark. The implementation details of the compared methods are similar to the previous subsection. As shown in Table 6, all instances of our framework are superior to their corresponding backbone networks by obvious margins. Moreover, our SANet+IADM and CSRNet+IADM outperform three advanced models (i.e., UCNet, HDFNet, and BBSNet) on all evaluation metrics. More importantly, our CSRNet+IADM achieves the lowest GAME(0) 4.38 and RMSE 7.05, and becomes the new state-of-the-art method on ShanghaiTechRGBD benchmark. This experiment shows that our approach is universal and effective for RGBD crowd counting.

6. Conclusion

In this work, we propose to incorporate optical and thermal information to estimate crowd counts in unconstrained scenarios. To this end, we introduce the first RGBT crowd counting benchmark with 2,030 pairs of RGB-thermal images and 138,389 annotated people. Moreover, we develop a cross-modal collaborative representation learning framework, which utilizes a tailor-designed Information Aggregation-Distribution Module to fully capture the complementary information of different modalities. Extensive experiments on two real-world benchmarks show the effectiveness and universality of the proposed method for multimodal (e.g., RGBT and RGBD) crowd counting.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No.U1811463, in part by National Key R&D Program of China under Grant No.2020AAA0109700, in part by National Natural Science Foundation of China under Grant No.61876045 and 61976250, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2017A030312006 and 2020B1515020048, in part by Major Project of Guangzhou Science and Technology of Collaborative Innovation and Industry under Grant 201605122151511.

References

[1] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *CVPR*, pages 4594–4603, 2020. 1

[2] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, pages 734–750, 2018. 3, 4, 5, 7

[3] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7. IEEE, 2008. 2

[4] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, pages 545–551, Sept 2009. 3

[5] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3, 2012. 3

[6] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016. 8

[7] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, 2020. 3, 7, 8

[8] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, pages 3052–3062, 2020. 3

[9] Min Fu, Pei Xu, Xudong Li, Qihe Liu, Mao Ye, and Ce Zhu. Fast crowd density estimation with convolutional neural networks. *EAAI*, 43:81–88, 2015. 3

[10] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. Cnn-based density estimation and crowd counting: A survey. *arXiv preprint arXiv:2003.12783*, 2020. 1

[11] Isha Ghodgaonkar, Subhankar Chakraborty, Vishnu Banna, Shane Allcroft, Mohammed Metwaly, Fischer Bordwell, Kohsuke Kimura, Xinxin Zhao, Abhinav Goel, Caleb Tung, et al. Analyzing worldwide social distancing through large-scale computer vision. *arXiv preprint arXiv:2008.12363*, 2020. 1

[12] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 423–431. Springer, 2015. 6

[13] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013. 3

[14] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, 2018. 2, 3, 8

[15] Bo Jiang, Zitai Zhou, Xiao Wang, and Jin Tang. cmsalgan: Rgb-d salient object detection with cross-view generative adversarial networks. *TMM*, 2020. 2

[16] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *CVPR*, pages 6133–6142, 2019. 3

[17] Di Kang, Debarun Dhar, and Antoni B Chan. Incorporating side information by adaptive convolution. In *NeurIPS*, pages 3870–3880, 2017. 3

[18] Di Kang, Zheng Ma, and Antoni B Chan. Beyond counting: comparisons of density maps for crowd analysis tasks—counting, detection, and tracking. *CSVT*, 29(5):1408–1422, 2018. 1

[19] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45, 2014. 3

[20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[21] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018. 1, 3, 4, 5, 7

[22] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *CVPR*, pages 1821–1830, 2019. 1, 2, 3, 4, 8

[23] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *CVPR*, pages 5197–5206, 2018. 8

[24] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Tianshui Chen, Guanbin Li, and Liang Lin. Efficient crowd counting via structured knowledge transfer. In *ACM MM*, 2020. 6

[25] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. *arXiv preprint arXiv:2007.08260*, 2020. 6

[26] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *ICCV*, pages 1774–1783, 2019. 1, 3

[27] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *IJCAI*, 2018. 1

[28] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. Dynamic spatial-temporal representation learning for traffic flow prediction. *TITS*, 2020. 1

[29] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *CVPR*, pages 3225–3234, 2019. 3

[30] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, pages 5099–5108, 2019. 1

[31] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Estimating people flows to better count them in crowded scenes. In *ECCV*, pages 723–740. Springer, 2020. 3

[32] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *ECCV*, 2020. 1

[33] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 3

[34] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019. 1, 3, 4, 5, 8

[35] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, 2020. 3, 7, 8

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8026–8037, 2019. 5

[37] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 2

[38] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, pages 9060–9069, 2020. 3

[39] Zhilin Qiu, Lingbo Liu, Guanbin Li, Qing Wang, Nong Xiao, and Liang Lin. Crowd counting via multi-view scale aggregation networks. In *ICME*, pages 1498–1503. IEEE, 2019. 3

[40] Weihong Ren, Di Kang, Yandong Tang, and Antoni B Chan. Fusing crowd density maps and visual object trackers for people tracking in crowd scenes. In *CVPR*, pages 5353–5362, 2018. 3

[41] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, volume 1, page 6, 2017. 3

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[43] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, pages 1879–1888. IEEE, 2017. 1, 3

[44] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, pages 1002–1012, 2019. 6

[45] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *TPAMI*, 2020. 2, 3

[46] Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, and Yin Wang. Leveraging crowdsourced gps data for road extraction from aerial imagery. In *CVPR*, pages 7509–7518, 2019. 2, 3

[47] Thirumalaisamy P Velavan and Christian G Meyer. The covid-19 epidemic. *Tropical medicine & international health*, 25(3):278, 2020. 1

[48] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *ECCV*, pages 660–676. Springer, 2016. 3

[49] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *ACM MM*, pages 1299–1302, 2015. 3

[50] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *TPAMI*, 2020. 2, 3

[51] Hao Wu, Hanyuan Zhang, Xinyu Zhang, Weiwei Sun, Baihua Zheng, and Yuning Jiang. Deepdualmapper: A gated fusion network for automatic map extraction using aerial images and trajectories, 2020. 3

[52] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, pages 5151–5159, 2017. 1, 3

[53] Lixian Yuan, Zhilin Qiu, Lingbo Liu, Hefeng Wu, Tianshui Chen, Pei Chen, and Liang Lin. Crowd counting via scale-communicative aggregation networks. *Neurocomputing*, 409:420–430, 2020. 3

[54] Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. *arXiv e-prints*, pages arXiv–2007, 2020. 2

[55] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *ICCV*, pages 6788–6797, 2019. 3

[56] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *ICCV*, pages 5714–5723, 2019. 1

[57] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015. 2, 3

[58] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, pages 8582–8591, 2020. 3, 7, 8

[59] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457. Springer, 2016. 1

[60] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *CVPR*, pages 8297–8306, 2019. 3, 7, 8

[61] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *CVPR*, pages 11046–11055, 2019. 5

[62] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and Jose MF Moura. Understanding traffic density from large-scale web camera data. In *CVPR*, pages 5898–5907, 2017. 1

[63] Shizhou Zhang, Yifei Yang, Peng Wang, Xiuwei Zhang, and Yanning Zhang. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *arXiv preprint arXiv:1910.11656*, 2019. 3

[64] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016. 1, 2, 4, 5, 6, 7

[65] He Zhao and Richard P Wildes. Spatiotemporal feature residual propagation for action prediction. In *ICCV*, pages 7003–7012, 2019. 5

[66] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2019. 3

[67] Desen Zhou and Qian He. Cascaded multi-task learning of head segmentation and density regression for rgbd crowd counting. *IEEE Access*, 2020. 2